Speaker-independent acoustic-phonetic recognition using adaptive vector quantization

Huan-Yu Su

INRIA

UNITE DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 940

Programme 5

SPEAKER-INDEPENDENT ACOUSTIC-PHONETIC RECOGNITION USING ADAPTIVE VECTOR QUANTIZATION

SU Huan-Yu

Décembre 1988

RR-0948

Speaker-Independent Acoustic–Phonetic Recognition using Adaptive Vector Quantization

Huan-Yu SU

Octobre 1988

Publication Interne n° 433

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

Speaker-Independent Acoustic-Phonetic Recognition using Adaptive Vector Quantization

SU Huan-yu[1]

Publication Interne n°433, 16 Pages, Octobre 1988

Abstract

Vector quantization has been employed to realize an acoustic-phonetic recognition system with success for continuous speech. The main idea is to build a dictionary by building first the sub-dictionaries corresponding to the phonetic classes, from a training set, using vector quantization; the recognition is then just a comparison of a vector with the elements of this dictionary. The accuracy of this system, for mono-speaker recognition, is about 70% and 90% using the nearest neighbor rule and the 3 nearest neighbors rule respectively. But this accuracy is getting unacceptable for a speaker-independent recognition (about 40% and 75% respectively), even with a pluri-speaker dictionary built from a 3-speaker training set. In order to realize an acceptable speaker-independent recognition system, two efficient algorithms allowing to adapt a dictionary to a new speaker, **in the process of recognition**, have been proposed and tested on the pluri-speaker dictionary. Also, a matching module has been wedged between the recognition module and the adaptation module to assure the quality of adaptation; the accuracy has been improved from 39% to 67% for the nearest neighbor rule and from 80% to 90% for the 3 nearest neighbors rule, after an adaptation in the process of 105 words' recognition (≈2 min speech).

Adaptation en cours de reconnaissance d'un dictionnaire de références phonétiques, à un nouveau locuteur

Résumé

La Quantification Vectorielle est utilisée pour réaliser un module de reconnaissance automatique de parole continue. L'idée principale est de construire un dictionnaire de références en réunissant des sous-dictionnaires, chacun correspondant à une classe phonétique; l'identification est de ce fait facilitée.

Une telle approche permet d'atteindre des taux de reconnaissance **phonétique** de l'ordre de 70 % et 90 % en employant respectivement la règle du plus proche voisin et des trois plus proches voisins, dans le cas mono-locuteur. Par contre, les performances se dégradent dans le cas indépendant du locuteur. Aussi nous envisageons d'adapter un dictionnaire pluri-locuteur (cad construit à partir d'un ensemble d'apprentissage pluri-locuteur), à tout nouveau locuteur, en cours de reconnaissance, sans nouvel apprentissage. Deux algorithmes sont étudiés : l'un est basé sur un algorithme de gradient stochastique, l'autre est une adaptation de l'algorithme de Lloyd. Afin d'améliorer réellement le dictionnaire de références, cet outil ne peut être utilisé que dans un système de reconnaissance complet ; le module lexical permet de corriger la plupart des erreurs de reconnaissance phonétique. Au cours de cette adaptation, le taux de reconnaissance **phonétique** (plus proche voisin) passe en moyenne de 40 % pour le dictionnaire initial, à 60 % après adaptation par le gradient stochastique, 67 % après adaptation par Lloyd, pour une prononciation de 105 mots. Quelques tests faits sur des enregistrements en salle machine nous permettent d'espérer en une certaine robustesse de ce type d'adaptation.

1. The author was with IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France, where this work, supported by the CNET-Lannion (Centre National d'Etudes des Télécommunications) under Contract CNET/INRIA N° 86 7B029007909245 LAA/TSS/DAP, has been realized.

I - Introduction

Speaker-independent continuous speech recognition is a problem especially difficult in comparison with the problem of isolated word recognition, since in the first case, we have to choose the recognition units, whereas in the second case the isolated word units are naturally identified. The beginning and the end of each unit are also difficult to determine in continuous speech unlike in the case of isolated word recognition. The recent approaches to continuous speech recognition all have one point in common: the first step is to perform the acoustic-phonetic decoding.

A vector quantizer (VQ) is a classifier which can find "the best way" to classify an ensemble of vectors. This technique has been widely used with good success in the areas of speech coding and isolated word recognition for several years [1,5,10,12,14,16].

In our approach, the vector quantization is used to realize a speaker-independent acoustic-phonetic recognition module for continuous speech. The main idea of this approach is to find a vector representation of speech signal, and to exploit this vectorial characteristic to build a ***dictionary*** from a training set. Then the recognition of a vector is just a comparison with the elements of the dictionary. It is clear that the reliability of the recognition depends essentially on the quality of the dictionary.

The way the dictionary is built may have an influence upon the quality of the dictionary: essentially, VQ partitions the representation space by minimizing a certain distortion measure, however, there are certain situations when this is not necessarily the best way of partitionning the space. For example, given an Euclidean space which consists of two sub-spaces of different vectors separated by a curved hyper-surface, and if we partition this space into two regions by minimizing the euclidean distortion measure, we can just find, at best, a hyper-plane as approximation of the curved hyper-surface. So, to avoid this eventuality of errors, our dictionary building consists of two stages: partition first the training set into sub-training sets, each corresponding to a phonetic class; then apply to each of them the algorithm of vector quantization to build a sub-dictionary. The dictionary is just the union of these sub-dictionaries.

We have realized a recognition in the mono-speaker case with very good reliability, but the attempt to use such a mono-speaker dictionary to realize a speaker-independent recognition system has proved to be difficult.

Much work was done recently towards solving such a problem of speaker-independence, particularly for systems which need a reference dictionary. In the case where vector quantization is employed, two principle approaches have been studied [5,6,10,13,15,17]:

- to build a multi-speaker dictionary; the training set is so large to get the dictionary capable to represent all kinds of speakers. We can also note that the module of vector quantization is just a pre-treatment before the recognition process as dynamic time-warping (DTW), hidden Markov models (HMM), ...

- to adapt the standard dictionary to the speaker by a very reduced training set; this idea is to find a couple of projectors towards a common space where the variability between speakers can be reduced. In this approach, even if the training set can be much reduced, the system is still necessary to be trained before the utilization.

Our approach is different from these mentioned before, since the adaptation of the dictionary to a speaker is ***dynamical***; this means the dictionary adaptation is performed in the process of recognition and a new static training set is no longer necessary. To realize this idea, we have first built a pluri-speaker dictionary (2 men and 1 woman) from a pluri-speaker training set which is the union of each mono-speaker training set, and then the dictionary is adapted to a new speaker in the process of recognition. This adaptation can be taken place by two adaptation algorithms:

1. stochastic gradient algorithm: the dictionary is modified at each identification in taking the barycenter between the element recognized and the entry vector.

2. generalized Lloyd algorithm: from the results of the recognition with the old dictionary, an automatic "training set" can be built, and then the generalized Lloyd algorithm permits to adapt the old dictionary to the speaker.

The first algorithm permits a permanent adaptation, whereas the second one is more efficient.

The organization of this paper is as follows: in **Section *II***, we describe the vector quantization algorithms used for dictionary building, and a mono-speaker recognition is realized and studied. We describe also the building of a pluri-speaker dictionary. In **Section *III***, we describe in detail two dictionary adaptation algorithms and their implementations, and the first results of dictionary adaptation are discussed. In **Section *IV***, we introduce the use of a lexical word recognizer as a matching module to realize a real dictionary adaptation. Finally, we summarize the results in **Section *V***.

II - Acoustic-phonetic recognition using VQ

A.Vector quantization and dictionary building algorithm

We focus our attention on a k-dimensional memoryless vector quantizer or static VQ, in a space E^k, which is just an operator Q mapping all vectors $\mathbf{x} \in E^k$ into a finite subset $\mathbf{D}$ of E^k, where $\mathbf{D}$ is called the dictionary of the VQ [8,9]. Now in supposing that each vector $\mathbf{x}$ represents a block of speech signal, and each sub-space $E_i \supset E^k$, defined by each element $\mathbf{y}_i \in \mathbf{D}$, corresponds to a phoneme, thus the recognition procedure consists of identifying this sub-space in which a particular representation vector lies.

The dictionary $\mathbf{D}$ of the recognizer can be built from a training set designed to be able to represent sufficiently the speech signal's representation space. In our case, this space E^k is the Euclidean space with $k=24$.

The goal of the dictionary building algorithm is to find the "best" partition of the training set's space. Since the initial dictionary is usually important, we use an algorithm permitting to get it reasonably, and the generalized Lloyd algorithm [9,11] is designed to find the "best" partition with an initial N-level dictionary given.

Our dictionary building algorithm [19] is a variation of the LBG algorithm which is the combination of the "splitting" algorithm and the generalized Lloyd algorithm [11].

(0) Initialization: Given a training set $\{\mathbf{x}_j, j = 1, \cdots, M\}$, and a distortion threshold σ;

(1) Define the centroid of the training set $\mathbf{y}_G$ as the initial dictionary $\mathbf{D}^1$

(2) Quantize the training space into the sub-spaces defined by the dictionary $\mathbf{D}^q = \{\mathbf{y}_i, i = 1,..., N_q\}$, and calculate the average distortion for each sub-space

$$dm_i = \frac{1}{\sum_{j=1}^{M} \delta(Q(\mathbf{x}_j), \mathbf{y}_i)} \sum_{j=1}^{M} d(\mathbf{x}_j, \mathbf{y}_i)\delta(Q(\mathbf{x}_j), \mathbf{y}_i);$$

where $\delta(Q(\mathbf{x}_j), \mathbf{y}_i(k)) = \begin{cases} 1 & \text{when } Q(\mathbf{x}_j) = \mathbf{y}_i(k) \\ 0 & \text{when } Q(\mathbf{x}_j) \neq \mathbf{y}_i(k). \end{cases}$

(3) Is $dm_i \leq \sigma$ for all i ? If so, set $\mathbf{D} = \mathbf{D}^q$ and halt. If not, "split" each vector $\mathbf{y}_i$ verifying $dm_i > \sigma$ into two close vectors $\mathbf{y}_i+\varepsilon$ and $\mathbf{y}_i-\varepsilon$, where ε is a fixed perturbation vector. The new dictionary $\tilde{\mathbf{D}}^{q+1} = \{\cdots, \mathbf{y}_i+\varepsilon, \mathbf{y}_i-\varepsilon, \cdots\}$ has N_{q+1} vectors (thus N_{q+1} is not always $2N_q$ as in the orignal LBG algorithm);

(4) The generalized Lloyd algorithm permits to find a best partition, in minimizing the average distortion, and a new dictionary $\mathbf{D}^{q+1}$. Increment $q = q + 1$ and return to step (2).

So in this version of the LBG algorithm, the cardinal of the dictionary $\mathbf{D}^{q+1}$ is not always two times of the cardinal of the dictionary $\mathbf{D}^q$. This permits to control easily the cardinal of the dictionary by varying the threshold σ (in speech recognition, no like speech coding as well, we do not need to fix the size of a reference dictionary).

B.Acoustic-phonetic recognition

The acoustic-phonetic recognition system using vector quantization has been realized in the mono-speaker case for simplifying the discussion. We have a database, 5 phonetically balanced lists[2] of 10 sentences, for each of 4 french speakers (3 males and 1 female). The signal is sampled at 12.8 kHz. A mono-speaker ***training set***, named as **L**, is gotten from 3 lists by a manual segmentation, and two other lists have been used to obtain a ***test set***, named **T**, by an automatic segmentation [2,4].

The problem of segmentation here is not to find the exact border between two

[2] · Each list has about 250 phonemes pronounced.

phonemes, but to place correctly an analysis window (Hamming window) on the signal to get a representative block for each phoneme. The Hamming window's width is 40 ms for the manual segmentation, and 40 or 20 ms for the automatic one, which gives the small segments < 40 ms (all segments whose width is < 20 ms are considered as transitions).

In the manual segmentation, if we can, that means the width of the phoneme is > 80 ms, there are three blocks taken for each phoneme: since the contextual influence is not the same for the beginning and the end of each phoneme, so a block is taken on each edge.

The test set is obtained by placing a Hamming window on the middle of each segment found by the automatic segmentation followed by an articulatory pre-treatment, which's two aims are to label first the segments corresponding to a transitional zone or a silence that we cannot get a correct recognition by our approach, and to determinate the explosion of /p,t,k,b,d, g / [3,4].

For each block obtained, a FFT has been taken and 24 Mel filters[3] have been applied to get a representation vector of E^{24}. In **Table 1**, we list the ensembles obtained for each speaker.

Table 1. For each speaker, the training set is obtained from 30 sentences (continuous speech) by a manual segmentation, and the test set is obtained from 20 sentences (different vocabulary from training set, continuous speech) by an automatic segmentation followed by an articulatory pre-treatment

Speaker	Training set	Test set
GM (male)	L_1: 1175 vectors	T_1: 569 vectors
CG (male)	L_2: 1255 vectors	T_2: 590 vectors
MG (female)	L_3: 1628 vectors	T_3: 564 vectors
JM (male)	L_4: 1363 vectors	T_4: 529 vectors

The dictionary building algorithm as we have introduced can just find a best way to partition the training space in finding the partition which minimizes the average distortion and get a corresponding dictionary. But if this partition does not coincide with the partition of the phonetic classes, not only we can not use directly this dictionary for the recognition with a good quality, but also the recognition process is more complicate [20]. So in our approach, the best partition has been determined to be coincident with the sub-spaces of phonetic classes which are supposed to be separated from each other. The training set is then partitioned into 21 sub-training sets corresponding to the 21 phonetic classes (of french langage) [18]:

3 · The scale has been determined by the CNET Lannion.

[a] = / a,ɑ/	[i] = / i /	[j] = / j /
[y] = / y,u /	[u] = / u /	[w] = / w /
[E] = / e, ɛ /	[&] = / ø,ə,œ /	[o] = / o, ɔ /
[on] = / ɔ̃ /	[an] = / ɑ̃ /	[un] = / œ̃ , ɛ̃ /
[m] = / m /	[n] = / n /	[bv] = / b,d,ɡ ,v /
[l] = / l /	[ge] = / ʒ /	[r] = / r /
[s] = / s /	[ch] = / ʃ /	[f] = / f /

and the dictionary building algorithm is applied to each of them to get a sub-dictionary corresponding. The dictionary is just the union of all these sub-dictionaries.

In our dictionary, there are not elements corresponding to some phonemes such as /p,t,k/, /ɲ/, /ŋ/, /z/, since our signal representation can not correctly represent the explosion of /p,t,k/, the same problem for the explosion of /b,d, ɡ/ which's voicing bar is coupled with /v/; and there are not enough vectors of /ɲ/, /ŋ/, /z/ in our training set to represent these phonetic classes. So in the test set, the blocks corresponding to these phonemes are eliminated.

We can wonder if some sub-spaces of phonetic classes intersect, and then the dictionary obtained cannot give good results in recognition. To evaluate the dictionary's quality, we have tested the accuracy of mono-speaker system on the training set **L** for each speaker, the recognition rates are over 90% and 99% using the nearest neighbor rule and the 3 nearest neighbors rule respectively. So this intersection problem of sub-spaces is not fundamental. In **Table 2** we find the recognition rates on the test sets and the corresponding confidence interval calculated with an error probability=0,05.

Table 2. Phonetic recognition rates using mono-speaker dictionaries

(with nearest neighbor rule)

Dictionary and number of elements / Test set	D_1	D_2	D_3	D_4
	140	198	215	218
T_1	75 ±4	52 ±4	32 ±4	38 ±4
T_2	55 ±4	70 ±4	44 ±4	51 ±4
T_3	38 ±4	43 ±4	65 ±4	42 ±4
T_4	40 ±4	38 ±4	40 ±4	54 ±4

(with 3 nearest neighbors rule)

Dictionary / Test set	D_1	D_2	D_3	D_4
T_1	94 ±2	80 ±3	64 ±4	73 ±4
T_2	84 ±3	94 ±2	71 ±4	79 ±3
T_3	63 ±4	67 ±4	88 ±3	67 ±4
T_4	66 ±4	65 ±4	65 ±4	82 ±3

The recognition results show us that, in the case of mono-speaker application, we can get good results, and since the dictionaries are smaller (≈190 elements) than a pluri-speaker dictionary, the recognition procedure is more rapide. But the recognition rates are not satisfying when a mono-speaker dictionary is used for speaker-independent recognition, especially when two speakers are of different sex.

In the second step, we have built a pluri-speaker dictionary from a pluri-speaker training set which is the union of three mono-speaker training sets $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{L}_3$ (two men and one woman). We holp that this pluri-speaker dictionary, having 496 elements, twice as big as a mono-speaker dictionary, represents better the representation space of all kinds of speakers than mono-speaker dictionaries can do.

The use of this pluri-speaker dictionary for speaker-independent recognition gives better results than in the case of using a mono-speaker dictionary, but these results are still not sufficiently accurate (for example, with $\mathbf{T}_4$, we have 43% and 74% as rates using the nearest neighbor rule and the 3 nearest neighbors rule respectively).

III - Dictionary adaptation algorithms

In this section, we discuss the problem of dictionary adaptation, whose aim is to modify the old dictionary dynamically, ***in the process of recognition***, with the new entries of the recognition system. This adaptation can be realized by either of two adaptation algorithms: the stochastic gradient algorithm or the generalized Lloyd algorithm.

A. Stochastic gradient algorithm [7,19]

This algorithm allows to "freshen up" a dictionary with a new ensemble. The operation's goal is, in modifying the old dictionary, to minimize the average distortion,

$$E\{ f(\mathbf{x} - \hat{\mathbf{x}}) \}$$

where $f(\mathbf{x} - \hat{\mathbf{x}})$ is a cost function, depending on the distortion measure. In the Euclidiean space, we may have

$$f_1(\mathbf{x} - \hat{\mathbf{x}}) = \frac{1}{2} \| \mathbf{x} - \hat{\mathbf{x}} \|^2$$

or $$f_2(\mathbf{x} - \hat{\mathbf{x}}) = \frac{1}{2} \| \mathbf{x} - \hat{\mathbf{x}} \|^2 \mathbf{1}_{\| \mathbf{x} - \hat{\mathbf{x}} \| \le R} + \frac{1}{2} R \| \mathbf{x} - \hat{\mathbf{x}} \| \mathbf{1}_{\| \mathbf{x} - \hat{\mathbf{x}} \| > R}$$

where R is a radius. Given a dictionary $\mathbf{D} = \{\mathbf{y}_i, i = 1, 2, ..., N\}$, we have

$$f(\mathbf{x} - \hat{\mathbf{x}}) = \min_i f(\mathbf{x} - \mathbf{y}_i)$$
$$= f(\mathbf{x} - \mathbf{y}_j)\, \delta(\mathbf{y}_j, \hat{\mathbf{x}}),$$

and the deviation

$$\frac{\partial}{\partial \mathbf{y}_j} f_1(\mathbf{x} - \hat{\mathbf{x}}) = (\mathbf{y}_j - \mathbf{x}) \delta(\mathbf{y}_j, \hat{\mathbf{x}})$$

or

$$\frac{\partial}{\partial y_j} f_2(x - \hat{x}) = \left[(y_j - x) \times 1_{\|x - y_j\| \le R} + \frac{y_j - x}{\|y_j - x\|} \times R \times 1_{\|x - y_j\| > R}\right] \times \delta(y_j, \hat{x})$$

gives the stochastic gradient algorithm as following: supposing that x_n is recognized as y_j^n of the dictionary D^n, thus $\hat{x}_n = y_j^n$, we can have a new dictionary D^{n+1} with just the j^{th} element y_j^{n+1} improved by:

$$y_j^{n+1} = y_j^n + \lambda(x_n - y_j^n)$$

or

$$y_j^{n+1} = y_j^n + \lambda(x_n - y_j^n) \times 1_{\|x_n - y_j^n\| \le R} + \lambda R \times 1_{\|x_n - y_j^n\| > R},$$

where λ is a step to choose, and

$$1_{\|x_n - y_j^n\| \le R} = \begin{cases} 1 & \text{if } \|x_n - y_j^n\| \le R \\ 0 & \text{if } \|x_n - y_j^n\| > R. \end{cases}$$

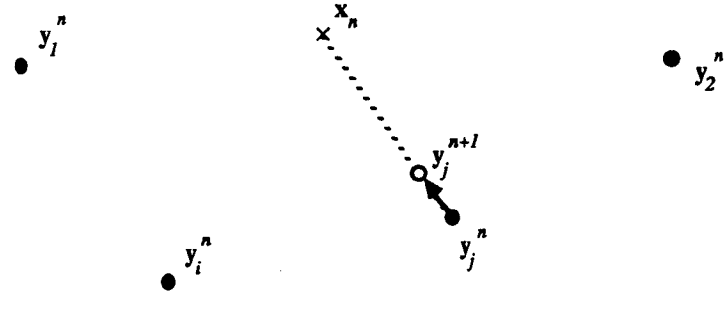

Figure 1. Dictionary adaptation by Stochastic gradient algorithm
At the n^{th} identification, the incoming vector x_n has been recognized as y_j^n, element of the dictionary D^n, then for the $n+1^{th}$ identification, the element y_j^{n+1} of the dictionary D^{n+1} is different from y_j^n.

This algorithm allows a permanent adaptation, but its reliability depends on the initial recognition rate using the nearest neighbor rule.

B. Generalized Lloyd algorithm [19]

From the results of the recognition, an "automatic training set" can be built: supposing that x_n is recognized as y_j, which is an element of the k^{th} sub-dictionary D_k of D, then x_n is memorized in the k^{th} sub-automatic training set. When the memory is filled up, the generalized Lloyd algorithm can be applied with this automatic training set to adapt the old dictionary to the new speaker.

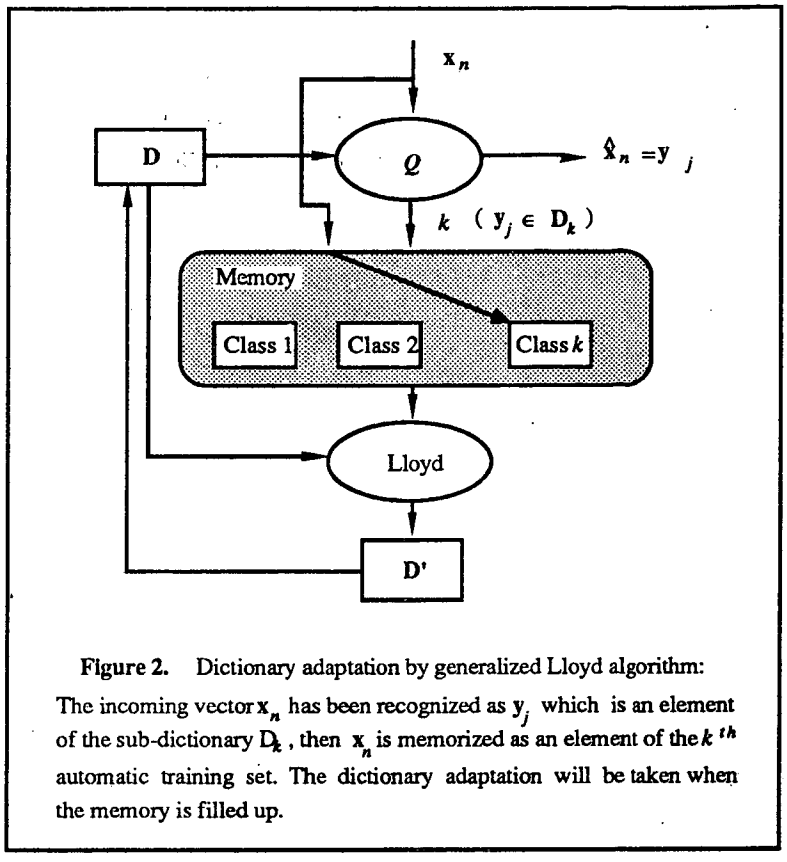

Figure 2. Dictionary adaptation by generalized Lloyd algorithm: The incoming vector x_n has been recognized as y_j which is an element of the sub-dictionary D_k, then x_n is memorized as an element of the k^{th} automatic training set. The dictionary adaptation will be taken when the memory is filled up.

This adaptation is more efficient, but the quality of the new dictionary **D'** depends on the quality of the automatic training set which depends itself on the initial recognition rate with the nearest neighbor rule. It is clear that the larger the automatic training set is, the better the quality of the adapted dictionary will be (but usually this size will be limitted by the memory capacity of the system).

C. Experiment

The first experiment has been made with the 4th speaker JM: the dictionary adaptation has been taken with the ensemble $\mathbf{L}_4$ (it was the training set of JM in our study for mono-speaker recognition, but now, it is just a manually segmented ensemble), and the quality of the new dictionaries is presented by the recognition rates with the test set $\mathbf{T}_4$.

Table 3. Recognition rates with the adapted dictionaries. The dictionary adaptation has been taken in the process of recognition (adaptation set: $\mathbf{L}_4$, test set: $\mathbf{T}_4$).

Adaptation algorithm / Rule	Lloyd	Stochastic gradient	
		f_1	f_2
Nearest neighbor	43	43	43
3 nearest neighbors	73	72	73

The results prove that, since the initial recognition rate, using the old pluri-speaker dictionary and the nearest neighbor rule, is not good enough (≈40%), the adaptation cannot improve the dictionary. Our next study intends to find a way to use the recognition results using 3 nearest neighbors rule.

IV - Dictionary adaptation with matching

In order to improve the quality of dictionary adaptation, we have to improve first the recognition rate by choosing one candidate from the 3 nearest neighbors. This job can just be done after a complete recognition of each word or sentence.

A. matching module

We have wedged a lexical module between our modules of phonetic recognition and dictionary adaptation. This lexical module is an isolated word recognizer based on a Hidden Markov Model, which recognizes a word in comparing each coming phonetic lattice with all references of its vocabulary, and finds out the word corresponding the most likely path by Viterbi algorithm. The matching is taken then with the phonetic lattice.

segment's number	nearest neigh-bor	2nd nearest neighb.	3rd nearest neighb.	phonetic reference matching	sub-dic. to adapt
1	..	..	..	..	--
2	**	**	**	k	--
3	o	on	w	o	[o]
4	m	n	l	n	[n]
5	y	&	E	E	[E]
6	..	..	..	..(k	--
7	..	..	..	..	--
8	**	**	**	t	--
9	&	E	r	&	[&]
10	r	s	l	r	[r]
11	s	f	bv	II	--
12	bv	m	n	bv	[bv]
13	E	bv	j	i	[i]
14	s	bv	&	z	--
15	i	j	l	i	[i]
16	bv	&	w	bv	[bv]
17	**	**	**	II	--
18	&	E	l	RR	[l]
19	l	bv	&	l	[l]
20	bv	l	E	II	--
21	bv	l	&	II	--
22	..	..	..	II	--
23	..	..	..	-/	--

where

(k
= omission of /k/
** = explosion of /p, t, k, b, d, /
RR= repetition (of [l] in our example)
I I = insertion
. . = silence (or transition in the phonetic lattice)
-/ = end of each reference

This matching module permits to improve the quality of dictionary adaptation in "correcting" the errors of the phonetic recognition, for example, the word "connecteur visible /kɔnɛktœ:r viziblʃ" has been recognized by the phonetic recognition module as before, and matched by the found phonetic reference [..-k-o-n-E-..-k-..-t-&-r-bv-i-z-i-bv-l], Then we can decide that:

-the 5th vector will be used to adapt the sub-dictionary corresponding to the phonetic class of [E] (but [y] is its nearest neighbor);

-the 14th vector will not be used to the adaptation since there is not a sub-dictionary corresponding to [z];

-the 11th ,17th, 20th and 21th will not either, since the matching module considers that they are insertions (they are perhaps transitions);

-the 13th will be used to adapt the [i]'s sub-dictionary even there is not [i] among its three nearest neighbors.

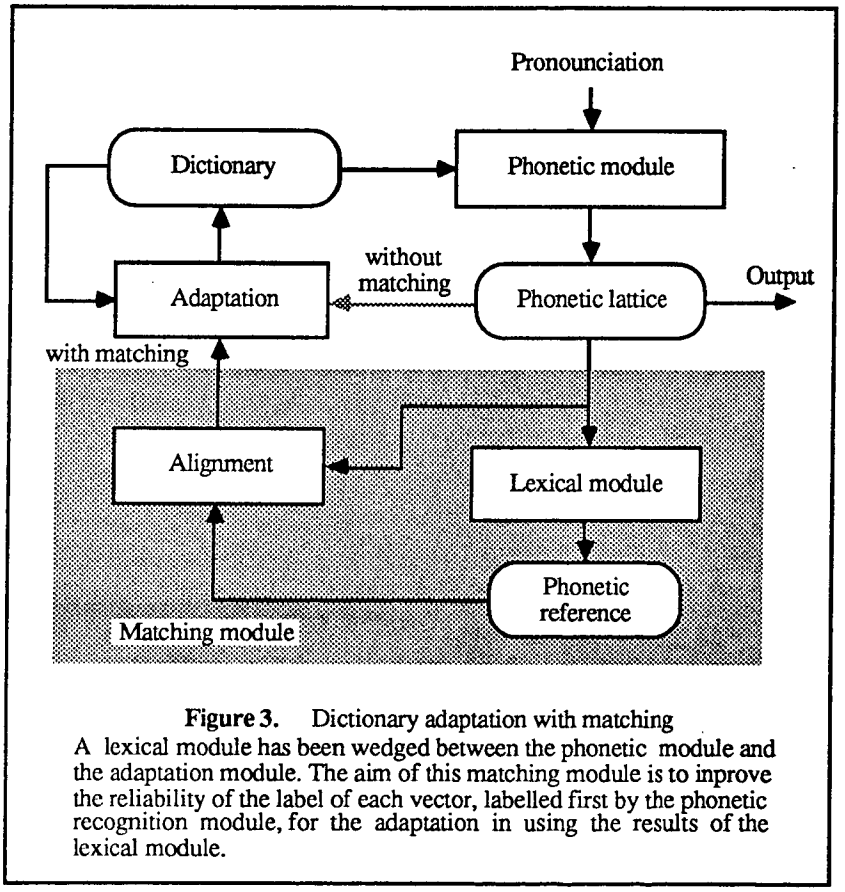

Figure 3. Dictionary adaptation with matching
A lexical module has been wedged between the phonetic module and the adaptation module. The aim of this matching module is to inprove the reliability of the label of each vector, labelled first by the phonetic recognition module, for the adaptation in using the results of the lexical module.

B. Experiment

Since the lexical module that we have used in our matching module is an isolated word recognizer, the experiment is, of course, taken with a database of isolated word: 10 pronounciations of 2 sentences (21 isolated words), by a new male speaker LM, the recording is performed in the same conditions as before and the signal is also sampled at 12.8 kHz.

The dictionary adaptation is taken in the process of 105 records' recognition: after the automatic segmentation, we have nearly 1200 segments, and the articulatory pre-treatment module eliminates nearly 300 of them (silences, explosions, some trantisions); the **word recognition** rate is about 90%, but an alignment is always taken even if a word is not correctly recognized; finally, the matching module gives 787 vectors to adapt the dictionary. The other 105 isolated words are used as test set (888 vectors obtained after segmentation and pre-treatment). We list the recognition rates using different dictionaries adapted by different algorithms in **Table 4.**

Table 4. Phonetic recognition rates with different dictionaries, the dictionary adaptation has been taken in the process of recognition. A matching module has been wedged between the phonetic recognition and the dictionary adaptation. The test set having 888 vectors is obtained from 105 isolated words of the same vocabulary as the adaptation set (the confidence interval is calculated with an error probability=0,05).

Dictionary / Rule	Initial pluri-speaker dictionary	Adapted by Lloyd	Adapted by stochastic gradient	
			f_1	f_2
Nearest neighbor	39 ±3	67 ±3	58 ±3	56 ±3
3 nearest neighbors	80 ±3	90 ±2	85 ±2	82 ±2

These results show clearly that the adapted dictionaries are much better than the initial one, and

-even though the generalized Lloyd algorithm is not sequential (since the dictionary cannot be adapted before an automatic training set has been built), it is a very efficient algorithm of dictionary adaptation;

-the stochastic gradient algorithm gives also the good results, especially the adaptation with f_1 seems better than which with f_2. These phenomenon explains perhaps that we do not need to limit the action of a "good" vector out of the matching module.

Our second experiment[4] has been performed with 4 new speakers different from all others mentioned before, two of them (speakers CL and VL) have pronounced those 2 sentences (isolated words) 20 times; and two others (speakers AC and AM) have pronounced those 2 sentences just 5 times (as set for adaptation) and 6 phonetically balanced lists of 10 sentences 2 times (as test set, continuous speech). The signal is still sampled at 12.8 kHz, but the difference is that the recording has been taken in a technic room without any precaution.

[4] · A sub-dictionary of [z] = / z / which has been added to the initial pluri-dictionary, is obtained from a sub-trainig set of / z / made with the french numbers (from 0 to 99) pronounced by 3 speakers (two males and one female) different from all speakers mentioned in this paper.

The dictionary adaptation for each speaker has been taken place in the process of recognition of 105 words as before (the rate of this word recognition is about 80% now). The other 315 words (for speakers CL and VL) and the 120 sentences (for speakers AC and AM) have been used to get a test set for each speaker. In **Table 5**, the size of each test set and the phonetic recognition rates using the initial dictionary and the adapted dictionaries are presented. The results prove that the adaptation improves always the quality of a dictionary and there is a very big difference between the initial dictionary and an adapted dictionary.

Table 5. Rates of phonetic recognition on test set of each speaker using the initial dictionary and the adapted dictionaries: The adaptation has been taken in the process of recognition of 105 isolated words for each speaker, with the generalized Lloyd algorithm and the stochastic gradient algorithm respectively. The test sets are, for CL and VL, 315 isolated words of the same vocabulary as the adaptation set, and for AC and AM, 120 sentences of continuous speech with different vocabulary (the confidence interval is ±2, calculated with an error probability=0,05)

Speaker	CL (male)				VL (female)				AC (male)				AM (female)			
Test set size	2289				2419				3284				3344			
Dictionary	Init.	Lloyd	Gradient		Init.	Lloyd	Gradient		Init.	Lloyd	Gradient		Init.	Lloyd	Gradient	
			f_1	f_2			f_1	f_2			f_1	f_2			f_1	f_2
Nearest neighbor	29	64	44	54	32	56	42	46	37	45	44	44	35	47	46	47
2 nearest neighbors	48	82	62	71	48	73	61	64	53	65	62	63	53	65	66	67
3 nearest neighbors	63	88	75	81	58	82	71	75	64	75	72	74	64	74	76	78

V - Summary

We have proposed a structure using the vector quantization to resolve the difficult problem of acoustic-phonetic recognition. Our results show that this approach gives very good reliability in the mono-speaker case, and dictionary adaptation is necessary in order to realize an acceptable speaker-independent recognition rate even with a pluri-speaker dictionary. This adaptation is realized by two proposed dictionary adaptation algorithms, both of which can improve the dictionary ***in the process of recognition***. The first results of this adaptation tell us that it is impossible to improve the dictionary if the initial recognition rate is very low (≈40%, using the nearest neighbor rule), but the dictionary is not damaged even with this low rate. Then we have proposed to wedge a matching module between the phonetic recognition and the dictionary adaptation, this realization gives very good results.

In our experimentation, the matching module is an isolated word recognizer, but it can be replaced by other modules which can, after a complete recognition of a unit (word or sentence), give the informations permitting to correct the errors of the phonetic recognition.

Our approach shows that the use of adaptative algorithms can help us to realize a speaker-independent recognition system without a very large training set which is always difficult to be obtained.

References

[1] J. P. ADOUL, J. L. DEBRAY, & D. DALLE : *"Spectral Distance measure applied to the optimum design of DPCM coders with L predictors"*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 512-515, April 1980

[2] R. ANDRE-OBRECHT : *"A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals"*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-36 n°1 pp. 29-40, January 1988

[3] R. ANDRE-OBRECHT et H.Y. SU : *"Expériences en vue du Décodage Acoustico-Phonétique à Partir d'une Recherche Statistique d'Evénements Articulatoires et d'un Codage Vectoriel"*. Proceedings, 16^{e} Journées d'Etudes sur la Parole, p.64-67, Hammamet, Tunisia, October 1987 (and will be publiched on Revue d'Acoustique Française, 1988)

[4] R. ANDRE-OBRECHT, B. DELYON, V. LE MAIRE, A. MORIN, H.Y. SU : *"Etude de Méthodes statiques pour le Décodage Acoustico-Phonétique de Parole Continue"*. Rapport de convention CNET/INRIA N° 86 7B029007909245 LAA/TSS/DAP, January 1988

[5] D.K. BURTON, J.E. SHORE, & J.T. BUCK : *"Isolated-word speech recognition using multisection VQ codebooks"*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-33 n°4 August 1985

[6] K. CHOUKRI and G. CHOLLET : *"Adaptation of Automatic Speech Recognizers to new speakers using cononical correlation analysis techniques"*. Computer Speech and Language, Vol.1 n°2, December 1986

[7] B. DELYON : *"Un théorème de limite centrale pour certaines équations différentielles aléatoires"*. Thèse de 3ème cycle, Université Pierre et Marie CURIE, Paris, France, July 1986

[8] A. GERSHO : *"On the structure of vector quantizers"*. IEEE Transactions on Information Theary, Vol. IT-28 n°2 , March 1982

[9] R. M. GRAY : *"Vector Quantization"*. IEEE ASSP Magazine vol.1 pp. 4-29, April 1984

[10] E. KOPEC, & M. A. BUSH : *"Network-based isolated digit recognition using vector quantization"*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-33 n°4 pp. 850-867, August 1985

[11] Y. LINDE, A. BUZO, & R. M. GRAY : *"An algorithm for vector quantizer design"*. IEEE Transactions on Communications COM-28 pp. 84-95, January 1980

[12] L. MICLET et M.DABOUZ : *"Un vocodeur à classification: transmission de parole à très faible débit par quantification vectorielle du spectre"*. « La Quantification Vectorielle pour le Traitement de la Parole », Actes du séminaire tenu à l'ENST, Paris, Febrary 1985

[13] K. C. PAN, F. K. SOONG, & L. R. RABINER : *"A vector-quantization-based preprocessor for speaker-independent isolated word recognition"*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-30 n°3, June 1985

[14] L. R. RABINER : *"A vector quantizer combining energy and LPC parameters and its application to isolated word recognition"*. AT & T Bell Laboratories Technical Journal Vol.63, n°5, May-June 1984

[15] L. R. RABINER, & S. E. LEVINSON : *"A speaker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building"*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-33 , June 1985

[16] M. J. SABIN & R. M. GRAY : *"Product code vector quantizers for waveform and voice coding"*. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, n°3, pp. 474-488, June 1984

[17] K. SHIKANO, K. F. LEE and R. REDDY : *"Speaker Adaptation through Vector Quantization"*. Internal rapport, Carnegie Mellon University, December 1986

[18] SU Huan-yu : *"Utilisation de la Quantification Vectorielle en reconnaissance de la Parole Continue"*. Proceedings, 15^{e} Journées d'Etudes sur la Parole, p.247-249, Aix-en-Provence, France, 1986

[19] SU Huan-yu : *"Reconnaissance Acoustico-Phonétique en Parole Continue par Quantification Vectorielle; Adaptation du Dictionnaire au Locuteur"*. Thèse de l'Université de Rennes I, November 1987

[20] D. VICARD, & L. MICLET : *"Steady part recognition of continuous speech for acoustic-phonetic decoding"*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.3 pp. 2263-2268, April 1986

LISTE DES DERNIERES PUBLICATIONS INTERNES

PI 429 **ESTIMATION DES SEGMENTS 2D : UN ALGORITHME ROBUSTE**
Ming XIE, Patrick RIVES
34 Pages, Septembre 1988.

PI 430 **UN ALGORITHME EFFICACE POUR LA MISE EN CORRESPONDANCE DES SEGMENTS 2D DANS UNE SEQUENCE D'IMAGES**
Ming XIE, Patrick RIVES
36 Pages, Septembre 1988.

PI 431 **TELEPILOTAGE DE ROV ASSISTE PAR ORDINATEUR**
Vincent RIGAUD, Lionel MARCE, Brigitte DUCHENE, Jean-Louis MICHEL
14 Pages, Octobre 1988.

PI 432 **PYRAMIDE, LOGICIEL DE MODELISATION GEOMETRIQUE POUR LA TELEOPERATION. MANUEL UTILISATEUR ET GUIDE DU PROGRAMMEUR**
Philippe EVEN
94 Pages, Octobre 1988.

PI 433 **SPEAKER INDEPENDENT ACOUSTIC-PHONETIC RECOGNITION USING ADAPTATIVE VECTOR QUANTIZATION**
Huan-Yu SU
16 Pages, Octobre 1988.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique